公共文化服务大数据集成架构设计研究*

■ 化柏林　赵东在　申泳国
北京大学信息管理系　北京 100871

摘　要：[目的/意义]针对当前各图书馆、文化馆等公共文化服务机构的多源异构数据，设计出一套行之有效的集成架构。[方法/过程]在充分分析公共文化大数据资源的基础上，对公共文化服务大数据的类型与分布进行分析，结合公共文化服务大数据的应用场景，设计公共文化大数据集成的架构。[结果/结论]提出一个由数据来源层、系统集成层、数据融合层、存储层、应用层五个层次构成的公共文化服务大数据集成架构，并对其中的采集、存储等关键技术进行研究。

关键词：图书馆　文化馆　公共文化　大数据　数据集成　集成架构
分类号：G250
DOI：10.13266/j.issn.0252-3116.2020.10.001

1　引言

我国图书馆主要有三大阵地，分别是公共图书馆、高校图书馆以及专业图书馆。在专业图书馆领域，于2000年成立了国家科技图书文献中心(national science and technology library, NSTL)，把科技领域的文献情报机构的资源整合起来，按照“统一采购、规范加工、联合上网、资源共享”的机制，采集、收藏和开发理、工、农、医各学科领域的科技文献资源，面向全国提供公益的、普惠的科技文献信息服务[1]。在高校图书馆领域，于1998年开始启动构建了中国高等教育文献保障系统(China Academic Library & Information System, CALIS)，建成以CALIS联机编目体系、CALIS文献发现与获取体系、CALIS协同服务体系和CALIS应用软件云服务平台等为主干，各省级共建共享数字图书馆平台、各高校数字图书馆系统为分支和叶节点的分布式“中国高等教育数字图书馆”，成员单位已有近两千家[2]。与高校馆与专业馆相比，公共图书馆面向普通大众，各地发展更加突出地方特色，这些原因使得公共图书馆领域一直没有一个全国性的统一平台。

近年来，随着国家对公共文化领域的重视，公共文化服务体系的建设已逐步开展起来，区域协同与跨馆服务的需求也日渐增强。很多城市已经实现了市一级和区一级的总分馆机制，在资源与服务上进行了集成与统一；上海等城市实现一卡通服务，市民一卡通除了交通、医疗挂号等服务以外，可以实现图书馆、文化馆等各类公共文化机构的统一认证与服务。国家公共文化云平台已经有上百家公共文化服务机构提供相应的资源、活动与服务。浙江嘉兴“文化有约”已上线多年，上海文化云的发展也非常迅速，“文化嘉定云”高度聚合区级图书馆、文化馆、博物馆、美术馆，以及街镇文体服务中心的文化资源和服务信息，通过网站、手机APP、微信、微博服务集群，为公众提供综合性、一站式、均等化的远程数字阅读、虚拟场馆体验、特色资源获取、文化活动预告、公共设施预订、线上交流展示等文化服务[3]。

随着公共文化服务在线平台的不断出现，公共文化服务数字资源数量也在不断地增长，在线平台上产生的数据也呈现出多样化趋势，对这些数据集成以后进行分析与挖掘，有着广泛的应用场景。从读者荐购到“你选书、我买单”，从借书排行榜、到馆统计等到大数据墙，从自助借还到机器人盘点上架，从网络点播到文化云、从馆际互借到文旅融合，这些都很好地把多源异构、甚至跨区域的资源数据、用户数据打通到一起并集成关联起来，为用户提供更好的精准服务，为公共文化服务机构提供实时动态的业务监测与管理，为管理

* 本文系公共文化服务大数据应用文化部重点实验室项目“公共文化大数据系统的集成应用研究”(项目编号:2017001)研究成果之一。
作者简介：化柏林(ORCID:0000-0001-9248-6455)，助理教授，博士，E-mail:huabolin@pku.edu.cn；赵东在(ORCID:0000-0001-5010-309X)，本科生；申泳国(ORCID:0000-0001-9157-5640)，硕士研究生。
收稿日期:2019-09-28　修回日期:2020-02-18　本文起止页码:3-11　本文责任编辑:徐健

部门提供更全面的现状揭示与决策支撑。而要实现这些应用,就需要把这些跨地域、跨机构、跨平台的数据集成到一起,这是公共文化服务大数据领域发展的前提与关键。

基于大数据集成技术解决公共文化领域里异构数据源数据之间的物理和逻辑层面的差异问题,给用户提供透明的一站式服务平台是公共文化服务大数据集成研究所要追求的目标。为了实现数据集成任务的目标,首先需要掌握好集成对象的数据特点,如数据类型、数据结构、数据量级、数据来源等;需要梳理并发现数据集成时会面临的问题,包括系统集成与数据融合时的问题;按照层次与流程设计公共文化服务大数据的集成架构,并对关键技术进行剖析。

2 相关研究述评

大数据在很多领域已实现用户画像与精准推荐、动态实时监测与远程监控、风险预警与趋势研判等。这些应用场景的实现,离不开各个领域不同数据源的数据集成。大数据集成与传统数据集成的不同点在于,从数据结构角度看,集成的数据对象不限于数据库里的结构化数据,还包括半结构化数据和非结构化数据,如日志数据、图像数据、视频数据、语音数据等。随着各类信息化系统的不断涌现以及数据收集的多样化,数据集成的问题在各行各业都已成为制约大数据挖掘利用的关键因素之一,大数据集成是大数据组织建设与分析挖掘的前提。

2.1 领域大数据集成应用研究

在电子商务领域,数据集成不仅有很多研究成果,而且通过数据中台等打通了所有的数据,包括用户注册数据、用户访问数据、交易数据、互联网金融数据、物流数据等等,通过数据集成与融合实现了各种应用。在地空遥感、农业生态、工业制造、智慧城市、图书情报等方面,大数据的集成也有一些研究与应用。

2.1.1 自然科学领域大数据集成

大数据集成起源于信号、遥感监测、工业自动化等领域。多源地理大数据为地理现象的分布格局、相互作用及动态演化提供了前所未有的社会感知手段[4]。王卷乐等提出了依托网络大数据、遥感大数据与社会经济大数据等地球大数据的集成与标准化框架,分析了网络数据获取与分析、遥感数据地表信息智能提取与处理以及社会经济数据空间化的关键技术[5]。赵芬等从数据获取、数据存储与管理、数据计算模式与系统和数据分析共 4 个模块详细阐述了生态环境大数据技术平台构建的关键技术[6]。吕佑龙等提出由物物互联层、对象感知层、数据分析层、业务应用层和云端服务层 5 个层次和一个大数据中心构成的智慧工厂技术体系架构[7]。李少波等认为大数据下制造业的五大关键技术,包括数据集成技术、数据存储技术、数据处理技术、数据分析技术以及数据展现技术[8]。王凇等认为未来的数据集成领域研究主要集中在对算法加速、对复杂数据源的集成以及基于众包的方法方面[9]。可以看得出,自然科学领域集成的数据源主要为传感器、遥感、遥测、卫星等硬件设备传输的数据,也称之为"硬数据"。数据集成是大数据分析与展现的前提与基础。

2.1.2 智慧城市大数据集成

智慧城市除了与像自然科学一样有一些来自传感设备的硬数据以外,也有一些管理和社会数据。政务信息领域由于数据集规模日益扩大,各部门的信息化过程不同,导致了各部门各层级之间信息不能充分地集成与共享,形成了信息孤岛问题。在面对这一现实问题,很多学者展开了探讨与研究。叶鑫等认为大数据与知识的"互联网 + 政务服务"云平台构建与服务,有助于消除信息孤岛、知识孤岛和业务孤岛[10]。杨兴凯等综述了在政府信息领域里所使用到的集成方法,认为针对电子政务信息资源整合标准化的研究比较少,导致电子政务标准化的数据模型和业务模型构建方面发展较缓慢[11]。洪之旭等提出一种分布式数据集成及可视化应用方法,基于大数据处理模式,将分散在不同网络路由的数据库数据接入、抽取和集成,进行挖掘分析,增强数据动态描述和 Web 可视化能力,提供面向服务的智慧化社会治理决策分析与应用[12]。刘岩等通过建设大数据中心实现异构数据源数据的集成,设计了以 Hadoop 为核心的异构数据源数据集成系统架构[13]。

2.1.3 情报大数据集成

情报机构与公共文化机构在数据资源、业务流程与服务功能等方面都具有很强的相似性,因此,情报领域的大数据集成对公共文化大数据集成也具有较强的参考借鉴意义。唐明伟等将数据集成的主要理论分为异构数据论和系统集成论,提出一个面向大数据的情报分析框架,大数据集群层是整个框架能够应对大数据应用的核心,主要由情报资源、计算集群和应用程序池三部分组成[14]。巴志超等通过对物理世界与人类社会中的元素或数据进行泛在协同感知与获取,将其映射到信息空间中实现数据的序化组织、信息融合与整合分析,进而反向指导人类社会与物理世界的决策

行为[15]。卢小宾等提出了一种通用的面向风险管理的银行大数据分析系统架构，旨在将不同类型的数据进行整合的基础上，构建统一、规范和易用的大数据分析系统[16]。陈伟等综述了海量异构数据集成、数据管理与分析方法和工具的开发进展，提出了建设数据驱动型科技情报研究模式的整体架构[17]。

2.2 公共文化大数据研究

在智慧城市、图书情报等人文社科领域，除了有些"硬数据"以外，还有加工数据、文档数据、社交关系数据等一些带有人工痕迹或社科属性的"软数据"，这些数据集成的流程、技术与方法具有很强的共通性。把多源的、异构的数据集成到统一的框架与平台下，可以更好地推动与促进公共文化大数据的发展与应用。图书馆为代表的公共文化服务机构有着丰富的数据资源，数据资源密集，而且很多数据是文本、视频等非结构化数据，具有大数据的典型特点，近年来围绕公共文化大数据的讨论也日渐增多。

2.2.1 公共文化大数据理论探讨

关于大数据与公共文化领域结合的探讨最早始于图书馆研究。在 2012 年，韩翠峰就意识到了大数据对图书馆功能的影响，指出了大数据将对图书馆的资源存储能力、用户需求挖掘能力等提出更高要求，需要图书馆改变技术开发与运用、数据集成与处理、人才培养与管理等方面的模式[18]。嵇婷等把公共文化大数据分为业务数据、网络数据、管理数据，探讨了公共文化大数据的采集、存储、分析方式[19]。苏新宁从资源建设、技术应用与服务三个方面展望了数字图书馆的未来发展[20]。刘炜等针对公共文化服务大数据发展的顶层设计，研究了这一过程中的政策与宏观管理、产业链与行业生态、技术标准规范等问题[21]。这些研究论证了大数据与公共文化服务结合的必要性，从不同视角对公共文化大数据进行了探索与剖析，建立了公共文化大数据应用的初步理论，为公共文化大数据应用方式的挖掘提供了理论支撑，对于大数据在公共文化服务领域的发展具有重要的指引与推动作用。

2.2.2 公共文化大数据体系研究

有了理论的指引，可以设计公共文化大数据体系。J. Li 等从人力资源、文献资源、技术支持、服务创新和基础设施构建五个方面论述了大数据在图书馆的应用框架[22]。曹树金等提出面向精准服务的图书馆大数据系统构建设想，系统结构包括多来源的数据采集层、数据预处理与存储层、精准化的数据分析建模层和支持精准化的管理与服务的应用层等自下而上的四个层级，系统的核心在于全面采集图书馆的大数据[23]。郭路生等基于 EA(企业架构)根据战略目标对应用体系的服务架构、IT 架构和治理架构对公共文化大数据应用体系进行顶层设计[24]。张春景将公共文化服务大数据应用模式分为三种驱动类型，包括数据驱动型、云平台驱动型和整体驱动型[25]。

2.2.3 公共文化大数据集成研究

除了这些体系的研究，有些学者专门提到数据集成或者技术平台的实现。李广建等认为公共文化服务大数据研究应着重关注公共文化服务大数据的概念与边界研究、方法研究、数据集成整合研究、用户画像建模研究、精准服务研究以及发展战略研究[26]。刘双等提出集成图书馆信息系统应由图书馆业务信息系统(library operating information system，LOIS)、图书馆管理信息系统(library management information system，LMIS)和图书馆服务信息系统(library services information system，LSIS)三者互联互通而成[27]。曹健等介绍了基于 Hadoop 的高校图书馆数字资源大数据分析系统，包括基础数据集成、读者标签化、资源分析、业务分析以及系统综合管理等五个功能模块[28]。图书馆的数据具有数据密集、非结构化数据分布广泛以及对服务的精准化诉求，使得图书馆大数据集成的问题日益迫切。

2.3 研究述评

从领域横向对比来看，在遥感监测、工业制造、农业生态、智慧城市等领域，大数据集成的研究已经比较充分而深入，这些研究成果可以为公共文化大数据集成提供参考与借鉴。与这些领域相比，公共文化大数据集成研究还刚刚起步。

从公共文化领域自身来看，公共文化领域对大数据的认识已比较充分，无论从业务发展、国家任务还是从用户需求来看，公共文化大数据的发展迎来了较好的发展机会与挑战，围绕大数据的研究也随之多起来，整体上来看，理论研究多一些，实践落地的研究还不够充分，另外，有多项研究提到了数据集成问题，但如何实现多源异构的数据集成，还缺乏专门的论述与探讨。因此，本文在充分分析公共文化大数据资源的基础上，结合公共文化服务大数据的应用场景，设计公共文化大数据集成的架构，并对其中所涉及到的关键技术进行剖析。

3 公共文化服务大数据资源分析

不同领域具有不同的数据资源，数据资源的分布

形态、数据结构、数据类型等决定着数据集成方式的选择。

3.1 公共文化服务大数据集成对象

3.1.1 数据来源

公共文化服务大数据集成对象的数据是图书馆、文化馆、博物馆、美术馆、纪念馆、群众艺术馆等服务机构所产生的。公共文化大数据的核心包括资源数据、用户数据、馆员数据、管理数据、服务数据、业务数据及其关系[29]。从公共文化大数据应用文旅部重点实验室的角度看,数据主要有开放数据、系统数据、基地加工数据和公共文化云数据。开放数据是指从图书馆、文化馆等服务机构以网络爬虫技术获取的服务数据以及年报中提取的业务数据。系统数据是指各文化服务机构的系统数据,主要存储在关系型数据库系统里,如 SQL Server、Oracle、MySQL、Sybase 等。基地加工数据为图书馆、文化馆、文化站和文化云的统计数据,这些数据以填报的方式或者以文件的形式传输。公共文化云数据包括基础数据、资源目录数据、资源内容数据、用户数据、活动数据等。

3.1.2 数据分类

由于来自各个数据源的数据结构以及数据所处理的模式并不相同,需要进一步将这些数据明确地区分是否属于结构化或半结构化或非结构化数据,以便明确哪些数据以何种数据采集技术和何种数据存储技术来处理。结构化数据具有明确且统一的数据结构,主要来自关系型数据库;半结构化数据一般带有一定的标记,且形成一定结构,例如以 XML 或 JSON 格式存储的数据;非结构化数据没有明确的结构,主要以文档、图片、音视频等文件形式存储。

系统数据是各服务机构提供的来自门户网站、管理系统、业务系统的关系型数据库的数据,其数据为结构化数据。文化云数据里的数据既有结构化数据又有非结构化数据,结构化数据包括文化云的基础数据、资源目录数据、用户基本数据和活动基本数据。半结构化数据主要指 XML 或 JSON 格式的日志数据,网络上带有标记的数据,如 MARC 数据、用元数据标记的文献题录数据,以及基地填报的数据,这些数据经过识别与转化后大部分内容可以转成结构化的数据,存入数据库,也可以以文件形式存储。非结构化的数据主要包括文化云上的活动通知文本、用户评论文本和资源内容等视频数据等,帖子、微博、微信等自媒体数据,还有各个机构网站上的 PDF 或 WORD 格式的论文、年报、研究报告等文档数据,以及其他网络自由文本数据。对公共文化服务集成对象数据的分类结构如表 1 所示:

表 1 公共文化服务大数据分类

结构化数据							半结构化数据			非结构化数据			
基地系统数据			文化云数据(非评论丨视频)				日志数据	网络数据	填报数据	文档数据	自媒体数据	网站数据	文化云数据
门户网站	管理系统	业务系统	基础数据	资源目录数据	用户基本数据	活动基本数据	XML JSON	网络上带有标记的数据	基地填报数据	年报 论文 研究报告	帖子 微博 微信	网络文本	评论文本 活动文本 音视频内容

3.2 数据集成面临的挑战

为了有效地解决数据集成问题,首先需要了解数据集成问题产生的原因。李亢等认为数据集成的难点主要可以归结为异构性问题、分布性问题和自治性问题。异构性问题主要是指各数据源的管理环境、数据模型、数据表达方式和数据语义的问题[30]。数据集成面临的问题主要包括系统集成问题与数据集成问题。

3.2.1 系统集成问题

建立公共文化服务大数据集成平台,需要实现不同数据源系统之间的无缝交流。即使不同数据源的系统都在同一个硬件平台上运行,并且全部使用支持 ODBC/JDBC 和 SQL 标准的关系数据库系统,也存在难以解决的问题。例如,虽然 SQL 为一种用于关系数据库的标准查询语言,但不同公共文化服务平台的实现方式有所差异,因此,在集成过程中需要对此差异进行协调。在数据集成中,集成的数据是来自已经存在于数据存储系统的数据,数据结构通常也比较复杂。此外,每个数据源提供的查询处理能力也大不相同。例如,一个数据源可能是支持完整的 SQL 的关系型数据库,因此,可容纳非常复杂的查询,但是数据源不只限于关系型数据库,也包含 WEB 和 TEXT、CSV、JSON 等文档数据源,对这些数据难以进行复杂的查询[31]。

3.2.2 数据集成语义问题

为相同目的建立同样的数据库,由于支持厂商不同,也可能设计出非常不同的数据模式,因此在数据语义、表达形式、数据源使用环境等多个方面呈现出异构性。多源数据异构性是数据集成面临的重要挑战,并且有效解决其异构性是保障数据集成质量的关键所在。这些问题包括语义歧义性、实例表示歧义性、数据不一致性、以及数据冗余、数据缺失等问题。

（1）语义歧义性。语义歧义性包括两个方面，有些数据用不同的名字来表示相同内容；也有相同的名字表示不同的含义。在不同公共文化服务机构基地加工数据库里对同一数据的描述可能不同，如图书馆和文化站的数据库里，对于到馆人次的字段名称描述不一致，图书馆基地加工数据将到馆人次表示为到馆人次，文化站将此描述为到站人次。为方便之后的统计应用需要采取合适的字段名来统一。不同数据源可能用相同的字段名来表示不同含义的字段。例如，当将网上咨询台系统与“一人一艺云平台”系统集成时，由于两个系统均有标题字段，但是它们的标题字段意义并不一致，网上咨询台系统标题字段指用户所咨询的标题，而“一人一艺云平台”标题指此平台发布的相关活动名称。此类问题可通过元数据映射来解决。

（2）实例表示歧义性。从各基地系统数据表来看，宁波市文化馆官网系统将用户的点击次数表示成点击率，但在数字资源访问系统里将此表示成点击次数。一个是表示成百分比，另一个是表示为次数，虽然两个数据类型同为数字型，但是还是在实例表示上存在歧义，即不同的数据源会用不同的方式描述同一实体，这就需要通过一定的转换规则对表示方式进行转换。时间与日期显示格式不一致也属于实例表示歧义性。这些数据通常是由系统日志而来，但是像服务机构网站上描述的活动数据时间与其网站数据库里记录的活动时间描述格式可能不同。不同来源的多种格式的同类数据，统一为某种格式即可。

（3）数据不一致性。造成数据不一致的原因有很多，包括同步问题、数据多分类、统计口径、计算错误、输入错误、过时的信息等。不同公共文化机构针对同一个实例的活动时间描述不一致、活动场所描述不一致，其中一个实例很可能是不准确的数据等。比如，年报里出现的大事记描述可能与对应的网页里内容描述不一致。此外，可能存在不同机构针对同一个讲座类别分类不同的问题，例如某个机构对某讲座归入文化类、然而另一个机构将此归入生活类。数据不一致性与实例表示歧义性的不同点是，数据不一致是因数据值的不同而产生，一般由于同步问题而发生，实例表示歧义性指的是同样的实例在表达形式上的不同。分布式大数据集成过程中数据属性特征在语义上的冲突特征包括：字符类型属性值的数据、数值类型属性值的数据、布尔类型属性值的数据，还有区间值类型属性值的数据四种[32]。针对现有关系数据库中分布式大数据集成冲突消解的问题可以划分成语义冲突、模式冲突以及实例冲突，其中语义冲突可以通过句法融合、逻辑树融合和频率融合法实现冲突消解[33]。

（4）数据冗余问题。公共文化服务大数据可分为数字馆藏资源数据和非馆藏资源数据，其中非馆藏资源数据分为讲座、展览和活动等的服务数据和参展、借阅、评论和投票等用户数据。由于公共文化服务机构非常多且对数据的理解程度与技术能力参差不齐，数据集成的过程中容易导致数据冗余、重复、错误。数据冗余指的是在同一个数据集上存在同样的数据。数据冗余问题包括三类：完全数据冗余、包含关系的数据冗余和部分数据冗余。完全数据冗余指的是要集成的异构数据源数据字段完全相同。包含关系的数据冗余问题指不同数据集上具有包含关系的数据。部分数据冗余指部分字段相同部分字段相异的情况。数据冗余问题一般通过取大舍小的方法来解决。

（5）数据缺失问题。数据缺失可能不是因为多源，而是人为错误、数据丢失、难以采集等造成的，数据可能不够完整或者同样的实例由于来自不同数据源而存在相对多余或缺少的属性字段。此情况下，在数据清理阶段可由以下的几种处理方法来解决：人工补填、全局常量填充、属性中心度量填充、最可能的值填充（回归、贝叶斯形式化方法或决策树归纳、忽略元组）[34]。

4　公共文化服务大数据集成设计

数据集成的主流模式，包括联邦数据库、数据仓库（数据复制架构）、中间件和基于本体的集成等4种模式。针对多源异构的数据特点，结合公共文化服务大数据领域特点，设计一个公共文化服务大数据集成架构，主要包括大数据集成的整体流程以及相关关键技术等。

4.1　公共文化服务大数据集成架构

对公共文化机构的各类数据进行全面的分析，调研相关的集成方法之后，形成公共文化服务大数据集成研究的解决方案，设计公共文化服务大数据集成架构。该架构分为5个层次：数据来源层、系统集成层、数据融合层、存储层，应用层，从流程上包括数据源获取、数据传输与采集、问题域分析、数据处理、数据存储以及数据应用等过程，具体内容见图1。

数据层主要包括4类数据来源，依次为开放数据、系统数据、基地的加工数据、公共文化云数据。按照数据的类型，分为实时数据、互联网数据、业务数据、日志数据等。

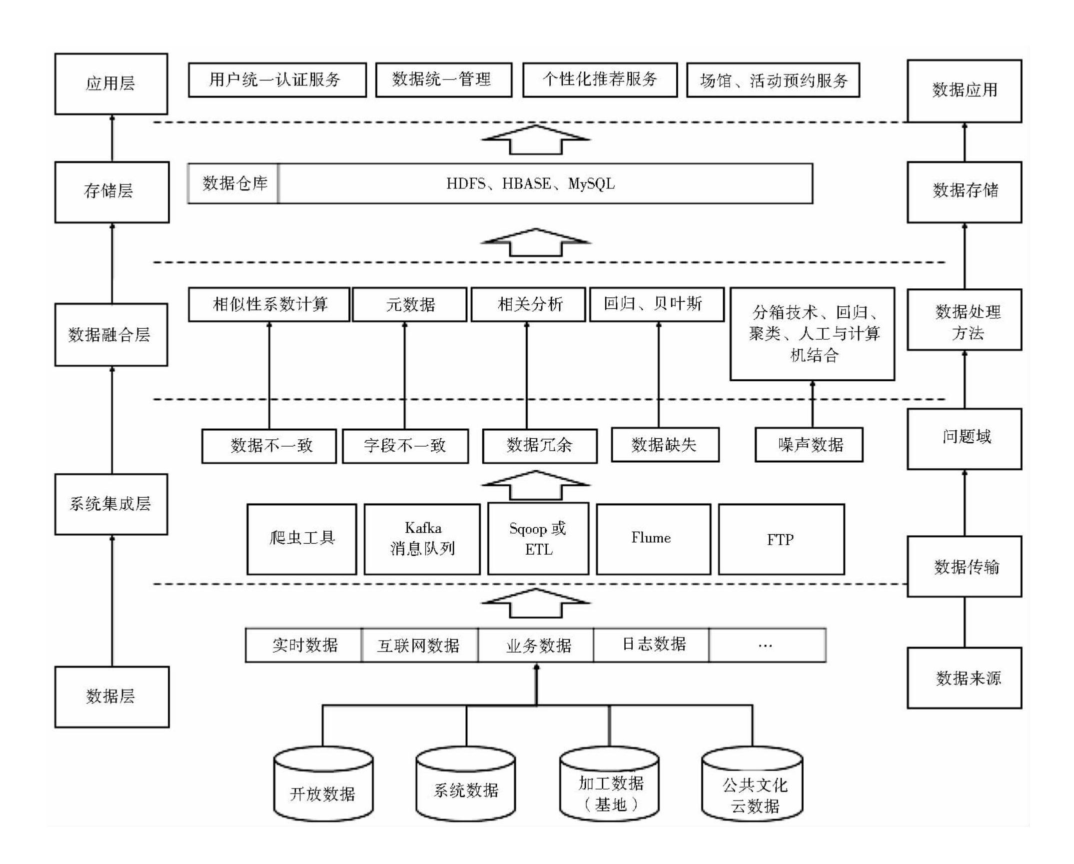


图 1　公共文化服务大数据集成架构

系统集成层主要负责数据的传输工作。在数据传输模块中借助不同的导入工具，实现不同元数据和不同结构数据的导入。其中，对实时性要求高的数据以分布式消息队列的形式由 Kafka 分发；关系型数据库使用 Sqoop 或 ETL 工具，直接将数据导入 HDFS 的数据库中；对于安全等级比较高的用户数据和一些离线数据，使用硬件复制或文件传输协议（FTP）传输的方式导入；对于日志等文本数据使用 Flume 工具导入；对于互联网数据使用爬虫程序爬取并导入。在数据集成的过程中会碰到数据不一致、字段名不一致、数据冗余、数据缺失、噪声数据等问题。

在数据融合层，主要由一些数据融合的方法支撑处理数据，包括相似性系数计算、元数据处理方法、相关分析、回归分析、贝叶斯判别、分箱技术、聚类技术、人工与计算机结合等多个处理方法。针对数据不一致问题，以相似性系数计算方法检测不一致的数据并将其统一成准确的数据；针对字段名不一致问题通过元数据技术统一处理为相同的字段名；针对不同数据冗余问题需要采取不同解决方法，在字段上出现的冗余问题可以由皮尔逊相关系数度量方法或者数据去重技术来解决，在图片上出现的冗余问题可以用数据压缩方法来解决；针对数据缺失问题可以由回归和贝叶斯方法来解决；针对噪声数据可以采用分箱技术、回归、聚类、人工与计算结合方法来解决。

在存储层，将已分类以及预处理完的数据根据特定需求分别存储在分布式文件系统 HDFS、分布式数据库 HBASE、关系型数据库 MySQL 中。使用分布式文件存储或非结构化 NOSQL 数据库进行存储，以保障上层高效地抽取数据。为提升数据分析的实时性和准确性，在计算层可采用适当的计算框架，像 Spark 的基于内存计算的开源集群计算系统或者像 Impala 的适用于大规模并行处理式 SQL 大数据分析引擎，可实现更快速的数据分析。

在应用层，基于解决系统集成与数据融合问题的前提下，可提供用户统一认证服务、数据统一管理、个性化推荐服务、场馆与活动预约服务、数字资源检索服务、文化资源服务和资讯发布管理等。

4.2　公共文化服务大数据集成分析关键技术

数据集成的前提是采集不同来源、不同结构的数据，集成的目的是为了深入分析与挖掘数据，以提高数据的应用价值。以 Hadoop 框架为底层的公共文化服务大数据集成技术框架，主要涉及到数据采集、数据存储、数据分析等关键技术。具体如图 2 所示：

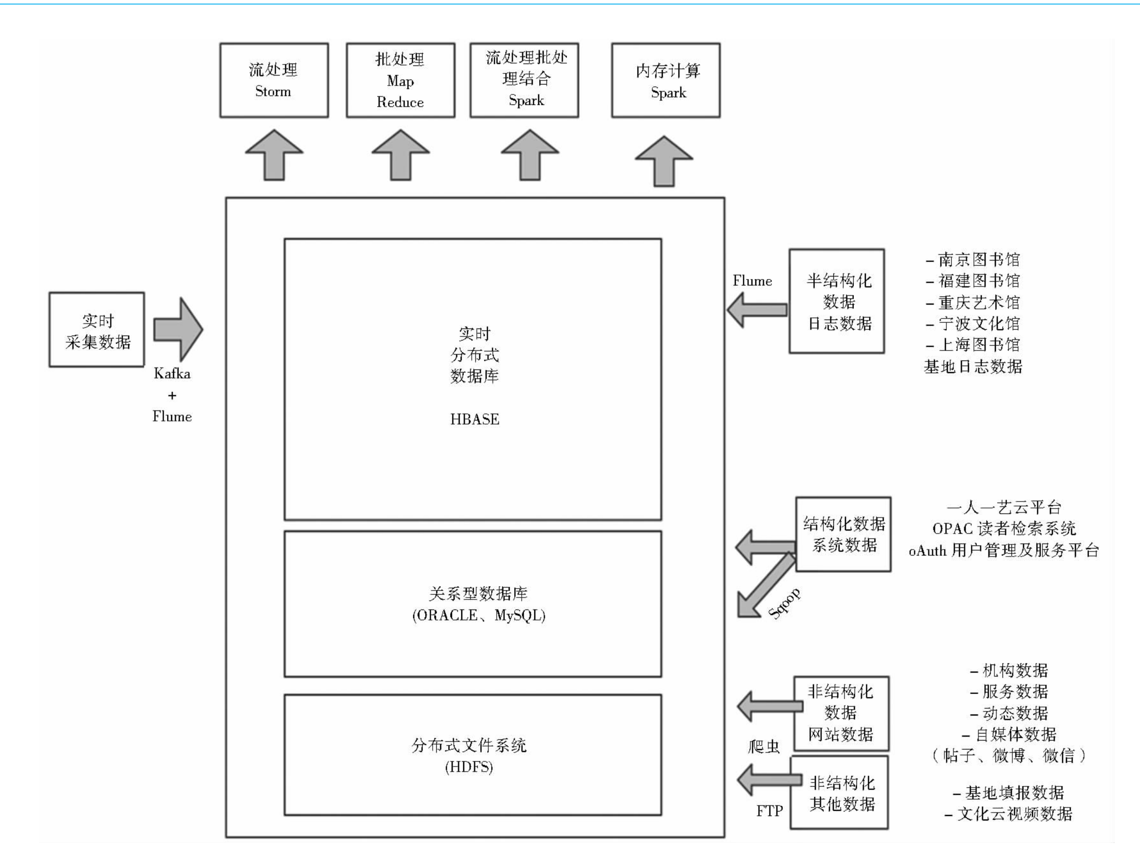


图 2　公共文化服务大数据集成技术框架

针对不同数据结构以适合的采集技术进行采集的流程。针对来自基地系统以及文化云的结构化数据，如“一人一艺云平台”、OPAC 读者检索系统、oAuth 用户管理及服务平台，可以用 Sqoop 工具进行数据采集，Sqoop 是用于结构化数据与 Hadoop 之间进行批量数据迁移的工具。针对来自南京图书馆、福建图书馆、重庆艺术馆、宁波文化馆、上海图书馆等各基地日志数据，通常为序列化半结构化数据，可以借助 Flume 进行采集。针对基地填报数据等的文档数据以 FTP 传输方式进行采集，现有 FTP 软件主要有 FlashFXP、FileZila、Cuteftp 等。针对机构数据、服务数据、动态数据、自媒体数据（帖子、微博、微信）等的网站数据可借助 Scrapy 等爬虫工具进行采集。

在采集层，针对在结构和时序上不同的各个数据源数据采用不同的采集工具进行数据抽取，对日志数据抽取可由 Flume 或 kafka 实施，也可以结合起来。对 MySQL、Oracle 等关系型数据库数据可由 Sqoop 工具来实施。对安全性要求较高的文档数据可由文件传输协议（FTP）采集。对从数据源已抽取的数据进行适当的转换和加工之后，可存储在 HDFS（分布式文件系统）上的 Hive（数据仓库）、HBASE（分布式数据库）和其他非关系型数据库中。

对于实时性较高的数据将 Kafka 与 Flume 两种技术结合利用，首先将业务数据实时存储到 Kafka 集群，然后通过 Flume 的 Source 组件实时处理 Kafka 的 Topic 获取的数据，将消费后的数据通过 Flume Sink 组件发送到 HDFS 或 HBASE 进行存储[35]。将两种技术结合使用的好处是借助 Kafka 工具可实时采集日志数据并以 Flume 高效地写入到 HDFS。

存储库包括分布式文件系统 HDFS、实时分布式数据库 HBASE、关系型数据库 ORACLE 和 MySQL。HBASE 适合存储海量半结构化数据，可以存储 Flume 工具采集的日志数据。ORACLE 和 MySQL 等关系型数据库，可存储结构化数据。由于 HDFS 适合存储半结构化数据或海量非结构化数据，因此可存储日志数据、以 FTP 传输方式采集到的文档数据以及以爬虫工具采集的网站数据。

数据处理单元部分除了提供基础的数据抽取与统计分析算法外，还提供半结构化和非结构化数据转结构化数据处理算法、数据内容深度理解算法等，涉及自然语言处理、视频图像内容理解、文本挖掘与分析等，数据处理效果的好坏直接决定了业务应用层数据统计分析的准确性和用户体验[13]。根据不同业务需求可以适当地选择使用数据处理技术，对于实时性要求高

的数据分析处理可选择使用流处理技术 Storm;对于进行大规模离线数据分析处理可使用批处理技术 Map - Reduce;对于整合流处理和批处理,实现数据的实时分析和深度挖掘可使用流处理与批处理相结合的技术 Spark;对于要求高性能的大数据分析处理能力也可利用基于内存计算的 Spark。

在数据源层上包含着集成对象,包括 MySQL、ORACLE、SQL server 等的关系型数据库,还有 XML 和 Excel 等的日志数据源和文本数据源。

在应用层,可得到的结果为基于已处理和融合的数据的分析结果,包括个性化推荐、日志分析、数据管理和用户统一认证服务等,实现实时监测、动态管理、精准服务以及决策支撑等。

5 结语

本文梳理了公共文化服务大数据集成中会面临的主要问题以及公共文化服务大数据集成架构。由于公共文化服务大数据结构并非只限于结构化数据,还包括图片、视频、评论等非结构化数据,传统数据集成方法难以实现集成。因此,笔者针对现有的大数据处理框架进行了简单的比较并设计了基于 Hadoop 框架的公共文化服务大数据集成架构。由于公共文化服务大数据来源多种多样,这些数据在结构和时序上都有所不同,需要采用合适的采集技术以及存储技术,由于各个数据源的数据模型不同,在数据融合层会面临一些逻辑上的问题,如数据不一致、字段不一致、数据冗余、数据缺失、噪声数据等,对这些问题的解决方案可以用相似性系数计算、元数据处理方法、相关分析、回归、贝叶斯、分箱技术、聚类技术、人工与计算机结合等多个处理方法。

本文根据描述的公共文化服务大数据集成中会遇到的问题,结合大数据集成相关技术设计出了公共文化服务大数据集成架构。当然,跨部门、跨机构的多源异构数据集成不仅仅是个技术问题,还有人员、管理与利益等方面的问题,大家并不愿意共享自己的数据,在智慧城市建设中尤为明显,公共文化领域亦是如此,大数据集成与融合还有很长的道路。当然,现在互联网公司在数据集成方面做得比较好,一方面这些企业获得了数据集成与应用所带来的巨大利益;另一方面,通过数据中台对数据进行集成、汇总,有了数据中台的支撑,可以针对新的环境变化、市场需求快速构建新的业务与系统,从而赢得竞争优势,这对公共文化领域是个启发。相信随着公众对公共文化服务需求的不断增长、各公共文化服务机构的不懈努力,公共文化服务大数据的集成也会逐步得到重视并有序地落地实现。

参考文献:

[1] 国家科技图书文献中心[EB/OL]. [2020 - 02 - 17]. https://www.nstl.gov.cn/Portal/zzjg_jgjj.html.

[2] 中国高等教育文献保障系统[EB/OL]. [2020 - 02 - 17]. http://www.calis.edu.cn/pages/list.html? id = 6e1b4169 - ddf5 - 4c3a - 841f - e74cea0579a0.

[3] 上海嘉定政府网[EB/OL]. [2020 - 02 - 17]. http://www.jiading.gov.cn/zwpd/zwdt/content_34571.

[4] 刘瑜,詹朝晖,朱递,等. 集成多源地理大数据感知城市空间分异格局[J]. 武汉大学学报(信息科学版),2018,43(3):327 - 335.

[5] 王卷乐,程凯,边玲玲,等. 面向 SDGs 和美丽中国评价的地球大数据集成框架与关键技术[J]. 遥感技术与应用,2018,33(5):775 - 783.

[6] 赵芬,张丽云,赵苗苗,等. 生态环境大数据平台架构和技术初探[J]. 生态学杂志,2017,36(3):824 - 832.

[7] 吕佑龙,张洁. 基于大数据的智慧工厂技术框架[J]. 计算机集成制造系统,2016,22(11):2691 - 2697.

[8] 李少波,陈永前. 大数据环境下制造业关键技术分析[J]. 电子技术应用,2017,43(2):18 - 21,25.

[9] 王淞,彭煜玮,兰海,等. 数据集成方法发展与展望[J]. 软件学报,2020,31(3):893 - 908.

[10] 叶鑫,董路安,宋禺. 基于大数据与知识的"互联网 + 政务服务"云平台的构建与服务策略研究[J]. 情报杂志,2018,37(2):154 - 160,153.

[11] 杨兴凯,刘畅. 政府信息资源集成方法研究综述[J]. 电子政务,2013(5):81 - 87.

[12] 洪之旭,陈浩,程亮. 基于大数据的社会治理数据集成及决策分析方法[J]. 清华大学学报(自然科学版),2017,57(3):264 - 269.

[13] 刘岩,王华,秦叶阳,等. 智慧城市多源异构大数据处理框架[J]. 大数据,2017,3(1):51 - 60.

[14] 唐明伟,苏新宁,肖连杰. 面向大数据的情报分析框架[J]. 情报学报,2018,37(5):467 - 476.

[15] 巴志超,李纲,安璐,等. 国家安全大数据综合信息集成:应用架构与实现路径[J]. 中国软科学,2018(7):9 - 20.

[16] 卢小宾,徐超. 面向风险管理的银行大数据分析系统架构研究[J]. 信息资源管理学报,2018,8(2):4 - 12.

[17] 陈伟,杨锐,何涛,等. 大数据环境下科技情报研究的新模式[J]. 科技导报,2018,36(16):78 - 85.

[18] 韩翠峰. 大数据带给图书馆的影响与挑战[J]. 图书与情报,2012(5):37 - 40.

[19] 嵇婷,吴政. 公共文化服务大数据的来源、采集与分析研究[J]. 图书馆建设,2015(11):21 - 24.

[20] 苏新宁. 大数据时代数字图书馆面临的机遇和挑战[J]. 中国图书馆学报,2015,41(6):4 - 12.

[21] 刘炜，张奇，张喆昱. 大数据创新公共文化服务研究[J]. 图书馆建设，2016(3):4－8.
[22] LI J,LU M，DOU G，et al. Big data application framework and its feasibility analysis in library[J]. Information discovery and delivery，2017：45(4):161－168.
[23] 曹树金,刘慧云,王连喜. 大数据驱动的图书馆精准服务研究[J]. 大学图书馆学报,2019,37(4):54－60.
[24] 郭路生,刘春年. 基于 EA 的公共文化服务大数据应用体系顶层设计研究[J]. 图书馆学研究,2019(5):31－37.
[25] 张春景,曹磊,曲蕴. 公共文化服务大数据应用模式与趋势研究[J]. 图书馆杂志,2015,34(12):4－8.
[26] 李广建,化柏林. 公共文化服务大数据研究的体系与内容[J]. 图书馆论坛,2018,38(7):62－71.
[27] 刘双,钱澄澄. 大数据集成图书馆信息系统体系框架与分类规范研究[J]. 图书馆学研究,2017,(5):31－37.
[28] 曹健,秦荣环,孙会清,等. 基于 Hadoop 的高校图书馆数字资源整合利用研究[J]. 图书馆工作与研究,2018,(3):74－78,101.
[29] 文庭孝. 大数据时代图书馆创新发展思考[J]. 图书馆,2019(5):15－22,27.
[30] 李亢,李新明,刘东. 多源异构装备数据集成研究综述[J]. 中国电子科学研究院学报,2015,10(2):162－168.
[31] DOAN A,HALEVY A,IVES Z. Principles of Data Integration[M]. 孟小峰，马如霞,马友忠,等译. 数据集成原理. 北京：机械工业出版社，2014.
[32] 梁勇. 关系数据库中分布式大数据集成冲突消解仿真[J]. 计算机仿真,2019,36(5):399－402.
[33] 王玥. 关系数据库中分布式大数据的集成冲突消解算法[J]. 科学技术与工程,2018,18(3):63－67.
[34] CSDN. 数据预处理[EB/OL]. [2020－02－17]. https://blog.csdn.net/weixin_42144636/article/details/81584372.
[35] CSDN. 使用 Flume 消费 Kafka 数据到 HDFS[EB/OL]. [2020－02－17]. https://www.cnblogs.com/smartloli/p/9984140.html.

作者贡献说明：
化柏林：研究思路与整体设计、综述部分撰写、论文修改；
赵东在：资料收集、部分论文撰写、图表绘制；
申泳国：资料收集、论文修改。

Research on Big Data Integration Architecture Design of Public Cultural Services

Hua Bolin　Cho Dongjae　Shin Youngkug
Department of Information Management，Peking University，Beijing 100871

Abstract：[**Purpose/significance**] To design an effective integration architecture for multi－source heterogeneous data of public cultural service institutions such as libraries and cultural centers. [**Method/process**] Based on the full analysis of the public cultural big data resources，this paper analyzed the types and distribution of the public cultural service big data，and designed the integration architecture of the public cultural big data in combination with the application scenarios of the public cultural service big data. [**Result/conclusion**] This paper proposes a public data service big data integration architecture consisting of 5 layers：data source layer，system integration layer，data fusion layer，storage layer and application layer，and discusses key technologies such as acquisition and storage.

Keywords：library　cultural center　public culture　big data　data integration　integrated architecture